



工程与应用

RFA-XGBoost模型在移动网络潜在投诉用户预测中的应用

张鹏, 高原

(中国移动通信集团设计院有限公司湖北分公司, 湖北 武汉 430021)

摘要: 为了提前预测并减少移动网络用户投诉事件的发生, 深入研究了多维数据分析在移动网络潜在投诉用户预测中的应用。采集移动网络用户广泛的业务域和运营域指标作为输入特征数据, 成功构建了基于极限梯度提升 (extreme gradient boosting, XGBoost) 算法的潜在投诉用户预测模型, 该模型在测试集上平均预测准确率达 96.35%。同时, 提出了迭代特征增强 XGBoost (recursive feature augmented XGBoost, RFA-XGBoost) 预测模型用于潜在投诉用户预测, 即通过不断迭代将前一轮 XGBoost 模型的预测输出作为新的特征添加到特征集中, 并重新训练新一轮的 XGBoost 模型, 优化后的平均预测准确率可提升至 98.89%。所提研究成果对于移动网络运营商而言, 意味着能够更早地识别并介入潜在投诉情况, 从而有效预防投诉事件的发生, 进一步提升用户满意度和服务质量, 具有重要的实践意义和商业价值。

关键词: 潜在投诉用户预测; 机器学习; XGBoost; 数据挖掘

中图分类号: TP391

文献标志码: A

doi: 10.11959/j.issn.1000-0801.2025036

Application of the RFA-XGBoost model in predicting potential complaint users in mobile network

ZHANG Peng, GAO Yuan

China Mobile Group Design Institute Co., Ltd., Hubei Branch, Wuhan 430021, China

Abstract: In order to predict and reduce the occurrence of complaints of mobile network users in advance, the application of multidimensional data analysis in the prediction of potential complaints of mobile network users was deeply studied. By collecting a wide range of business domain and operation domain indicators of mobile network users as input feature data, a potential complaint user prediction model based on extreme gradient boosting (XGBoost) was successfully constructed, which had an average prediction accuracy of 96.35% on the test set. At the same time, the recursive feature augmented XGBoost (RFA-XGBoost) prediction model was proposed for the prediction of potential complaint users. By iteratively adding the predicted output of the previous round of XGBoost model to the feature set as a new feature and retraining the new round of XGBoost model, the average prediction accuracy after optimization could be improved to 98.89%. For mobile network operators, the research results mean that they can identify and in-



tervene in potential complaints earlier, so as to effectively prevent the occurrence of complaints and further improve user satisfaction and service quality, which has important practical significance and commercial value.

Key words: prediction of potential complaint users, machine learning, XGBoost, data mining

0 引言

随着移动互联网的飞速发展,移动网络已经成为人们日常生活中不可或缺的一部分。当前,随着网络规模的扩大和用户需求的多样化,移动网络运营商面临着诸多挑战,其中用户投诉问题尤为突出^[1]。用户对移动互联网产品的关注点逐渐从价格转移到质量上,维权意识的提高和信息反馈渠道的便利,使得运营商投诉量激增。因此,运营商把网络用户满意度作为服务质量反馈的一个重要指标。传统的网络质量监控和投诉处理方法往往被动响应,难以提前预防^[2]。因此,对移动网络潜在的投诉用户进行主动预测以避免投诉产生,成为提升用户服务质量和优化网络运营的重要课题^[3-6]。

以下是关于移动网络潜在投诉用户主动预测研究现状的简要介绍。陈秀敏等^[7]提出了基于数据挖掘的4G用户投诉预测方法,用实际运营域数据特征进行建模,并对比了随机森林、逻辑回归等若干种机器学习算法的预测精度,其中随机森林的预测综合评价指标最高。但事实上,移动网络用户对网络使用情况的满意度评价是一个复杂的问题,因为它受到多种客观和主观因素的共同影响。除了受到包括无线网络覆盖情况、网络稳定性等运营域客观因素的影响,还会受到用户偏好、用户属性等众多业务域主观因素的影响。随着人工智能领域机器学习算法的发展,目前已有诸多机器学习算法的效果和效率显著优于随机森林。宗宇雷等^[8]提出了一种基于极限梯度提升(extreme gradient boosting, XGBoost)算法^[9]的投诉用户特征识别和预测方法,通过收集投诉用户的主客观数据特征,并引入XGBoost算法预测

潜在的投诉用户,提早介入网络优化和用户关怀,有效地防止网络投诉的发生。然而,该方案所使用的用户主观数据过于单一,只采用了用户的性别、年龄等6个指标作为用户主观数据特征。王可争等^[10]提出一种利用反向传播(back propagation, BP)神经网络预测用户投诉的方法。然而,其实验中采用的用户特征维数和用户样本量均极小,限制了模型的使用场景。这是由于BP神经网络本身的局限性:在过高的用户特征维度和用户样本量下,模型训练和推理的效率极为低下。马洁等^[11]结合运营域网络质量和业务域用户特征等关键指标数据进行决策树模型建模与预测,但是决策树模型泛化性较差,易产生过拟合现象。周文杰等^[12]提出了一种基于深度信念网络(deep belief network, DBN)模型^[13]的用户投诉预测模型,学习构造有效的高层非线性组合特征,使得预测模型在以往模型的基础上提升了7.1%。然而,DBN作为深度学习模型,包含了多个层次和大量的参数,其训练过程通常比传统机器学习算法如XGBoost、轻量级梯度提升^[14](light gradient boosting machine, lightGBM)、自适应提升^[15](adaptive boosting, AdaBoost)、梯度提升决策树^[16](gradient boosting decision tree, GBDT)等更为复杂和耗时,并不能满足大规模移动网络投诉用户预测模型的快速迭代需求。

综上,本文旨在构建一种综合多维数据来进行移动网络潜在投诉用户快速、准确预测的方法,即通过用户终端、业务使用行为、常驻区域网络情况、套餐、用户属性等主观和客观多维信息,预测用户是否为潜在投诉用户。移动网络投诉用户预测流程如图1所示。首先,对网络用户

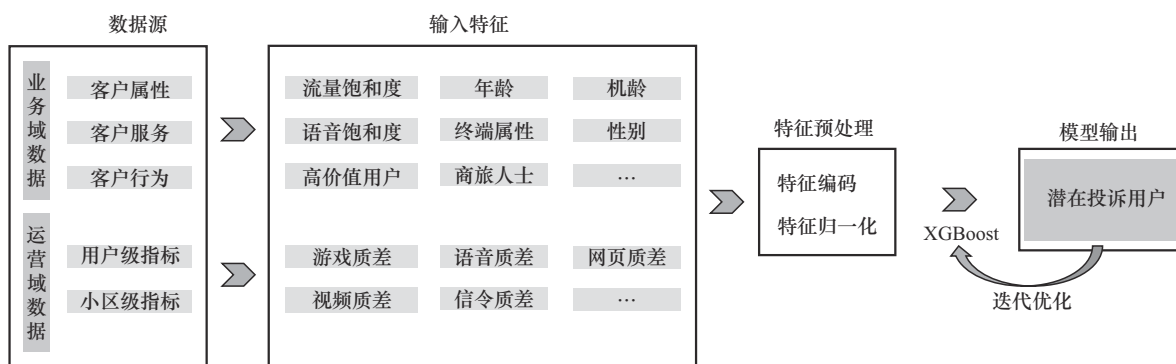


图1 移动网络投诉用户预测流程

的业务域数据（如客户属性、客户服务、客户行为等）和运营域数据（如用户级指标、小区级指标等）等多项指标进行采集，经过数据预处理之后作为输入特征，将用户是否投诉的回复结果作为标签数据。之后，建立基于XGBoost的潜在投诉用户预测模型。模型输入为经过数据预处理及样本均衡后的用户特征数据，模型输出其投诉与否的分类结果。除此之外，本文还创新性地提出迭代特征增强XGBoost（recursive feature augmented XGBoost, RFA-XGBoost）预测模型，即通过不断迭代地将前一轮XGBoost模型的预测输出作为新的特征添加到特征集中并重新训练新一轮的XGBoost模型，以不断提升预测准确度。在预测出潜在投诉用户后，可以对其进行精准改善与高效资源关怀，以提升移动网络用户满意度，增强市场竞争力。

1 数据准备

1.1 数据采集

为进行投诉用户预测建模，需要划定样本用户范围，并获取样本用户的相关原始数据，包括运营域的26个指标和业务域的48个指标。需要采集的用户数据见表1。

表1列举了包括运营域和业务域共74项数据指标，采集后共同作为输入用户特征。其中，某些指标如果一天内在不同时段变化明显，可以按

每2~4 h取均值后进行采集。除此之外，需要对上述样本用户进行短信或电话调研，同时收集反馈的投诉与否的结果，作为模型训练的标签，未投诉标记为0，投诉标记为1。对某省部分移动网络用户进行调研后，收集到的数据集中用户分布见表2。

1.2 数据预处理

由于表1中的数据指标较多，数据收集过程中不可避免存在某些用户的个别指标数据缺省或出现异常值情况。传统的数据处理方法可能需要花费大量的时间和精力来清洗和预处理这些数据，而XGBoost等树模型能够自动处理缺省值，从而在一定程度上减少了数据预处理的复杂性和时间成本。因此，这里不对缺省值或异常值做额外处理。

除此之外，表1中的指标不仅包括数值型特征（数据质差次数、语音失败次数等），也包括分类型特征（是否宽带用户、是否高校客户、是否高价值用户等）。数值型特征需要进行数据归一化，将数值规整到同一动态区间 $[-1, 1]$ ，提升模型收敛效果，避免过拟合；分类型数据需要进行独热分类编码，以便进行统一建模与分析。

将原始数据集分为3部分，分别为训练集（80%）、验证集（10%）和测试集（10%）。划分好训练集、验证集和测试集即已经完成数据准备工作。



表1 需要采集的用户数据

指标维度	指标名称	指标维度	指标名称
运营域 特征	数据质差次数	业务域 特征	是否商旅人士
	数据单据次数		宽带当月活跃天数
	数据质差占比		近3个月的投诉次数
	信令质差次数		折扣金额
	信令次数		近3个月话费查询次数
	信令质差占比		是否服务投诉
	语音失败次数		是否营销投诉
	语音次数		月接通电话通数
	语音失败占比		用户信用度值
	总质差次数		是否家庭宽带用户
	总单据次数		5G开关状态
	总质差占比		真实消费金额
	视频单据数		是否5G网络用户
	视频传输控制协议（transmission control protocol, TCP）第3次握手时长质差次数		5G未登网用户
	视频TCP第3次握手时长质差占比		流量饱和度
	视频下行往返时延（round-trip time, RTT）质差次数		客户信用度
	视频下行RTT质差占比		是否VoLTE活跃用户
	视频关键质量指标（key quality indicator, KQI）质差次数		是否语音超套
	视频KQI质差占比		是否资费投诉客户
业务域 特征	网页单据数		是否4G+手机终端客户
	网页TCP第3次握手时长质差次数		性别
	网页TCP第3次握手时长质差占比		互联网电视当月活跃天数
	网页下行RTT质差次数		是否高价值用户
	网页下行RTT质差占比		是否不限量限速用户
	是否不限量用户		是否中国移动微信公众号关注客户
	是否宽带用户		近3个月10086拨打次数（含速拨号）
	年龄		日均语音计费时长
	终端是否支持长期演进语音承载（voice over long-term evolution, VoLTE）		是否宽带投诉客户
	是否开通VoLTE业务		是否流量超出套餐上限
	当月平均用户收入（average revenue per user, ARPU）折后（元）		是否员工卡
	当月每用户平均数据使用量（data usage per user, DOU）（MB）		是否移动副卡异网主卡用户
	终端是否支持5G		是否免打扰
	5G是否激活		是否高校客户
	日均语音话单数		是否费用质疑投诉
	语音饱和度		是否投诉常客
	机龄（月）		月通信天数

表2 数据集中用户分布

反馈结果	样本数	样本比例
非投诉用户	235 396	82.79%
投诉用户	48 939	17.21%

2 基于 XGBoost 的移动网络投诉用户识别模型构建

XGBoost 是一种在梯度提升框架下实现的高效、灵活且可扩展的机器学习算法。其基本原理是将多个弱分类器的结果结合起来，形成一个强分类器。这些弱分类器是有先后顺序的，后一个弱分类器的训练会基于前一个弱分类器的预测结果，使得先前分类器做错的样本在后续受到更多的关注。XGBoost 具有较高的模型预测精度，也能够快速迭代。由于本任务中涉及的用户数量极大、涉及的特征维度也较高，因此选择 XGBoost 可以在该大规模数据集中做模型的快速训练、推理以及更新迭代。

在经过数据预处理之后，得到了初始化用户特征集 x_i 和标签 y_i ($i=1, 2, \dots, N$)，其中 N 为样本量。该算法的具体步骤如下。

(1) 设置初始模型 $f_0(x)$ 。

(2) 迭代计算，迭代次数为 $t=1, 2, \dots, T$ 。

① 计算当前模型 $f_{t-1}(x)$ 在每个样本点 x_i 上的预测值 $\hat{y}_i^{(t-1)}$ 和实际值 y_i 之间的残差。

② 训练一棵新的决策树 $f_t(x)$ ，使其能够最小化目标函数 $J^{(t)}$ ：

$$J^{(t)} = \sum_{i=1}^n [L(y_i, \hat{y}_i^{(t-1)} + f_t(x_i))] + \Omega(f_t) \quad (1)$$

其中， L 是损失函数， Ω 是正则化项。

③ 为了简化计算，使用泰勒二阶展开式对目标函数进行近似：

$$J^{(t)} \approx \sum_{i=1}^N [L(y_i, \hat{y}_i^{(t-1)} + g_i f_t(x_i)) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \quad (2)$$

其中， g_i 和 h_i 分别是损失函数 L 在 $\hat{y}_i^{(t-1)}$ 处的一阶

导数和二阶导数。

④ 将树的结构（叶子节点和分裂点）和叶子节点的输出值作为参数，将目标函数参数化为：

$$J^{(t)} = \sum_{j=1}^T \left[G_j w_j + \frac{1}{2} (H_j + \lambda) w_j^2 \right] + \gamma T \quad (3)$$

其中， T 是叶子节点的数量， w_j 是第 j 个叶子节点的输出值， G_j 和 H_j 分别是分配到第 j 个叶子节点的样本的一阶导数之和及二阶导数之和， λ 和 γ 是正则化项的系数。

⑤ 通过优化算法，根据参数化后的目标函数构建最优的树结构。在构建过程中，会不断选择最优的特征和切分点，以最小化目标函数。

(3) 得到最终模型。将 T 个决策树的结果相加，得到最终的预测值：

$$\hat{y}_i = \sum_{k=1}^T f_k(x_i) \quad (4)$$

从式 (4) 可以看出，最终模型可以用于对新样本 x_i 进行分类或者回归预测，通过计算 $\hat{y}_i$ 的值来确定样本 x_i 最可能的投诉产生与否的预测结果。当其被指定为回归任务时，预测值可以被认为划分为投诉用户或者非投诉用户的概率值。该值越接近 1，表示模型更倾向于判定该用户为投诉用户；反之，越接近 0 则表示模型更倾向于判定该用户为非投诉用户。而当其被指定为分类任务时，预测值即划分为投诉用户或者非投诉用户的分类结果。

3 基于 RFA-XGBoost 的移动网络投诉用户识别模型构建

RFA-XGBoost 算法是本文创新性提出的一种集成学习方法。其核心在于，通过不断迭代地将前一轮 XGBoost 模型的预测输出作为新的特征添加到特征集中，使得后续轮次的模型能够在更丰富的特征空间上进行学习。这种方法不仅利用了 XGBoost 在处理高维数据和复杂非线性关系方面的优势，还通过迭代增强特征集的方式，逐步提



升了模型对特征集的理解和预测能力。

RFA-XGBoost 算法的训练和测试流程如图 2 所示。其具体流程步骤如下。

(1) 将第 2 节中基于 XGBoost 模型的移动网络投诉用户识别模型作为原始模型 (XGBoost-0), 在原始训练集上开展模型训练, 损失函数可以定义为回归或者分类损失函数。

(2) 原始模型训练完成后, 将其在训练集上的预测结果作为新的特征, 与原始 74 维训练集合并, 形成新的特征训练集, 特征维度变为 75。

(3) 使用更新后的训练集迭代训练新的 XGBoost 模型 (XGBoost-1)。

(4) 重复上述步骤 $n-1$ 次, 获得每次迭代下的模型权重, 并在每次迭代时都将前一轮训练集的预测结果作为新特征添加到特征训练集中, 测试集的预测结果作为新特征添加到测试集中, 特征维度不断扩充。

(5) 直达到达预定的迭代次数 n , 此时特征维度为 $74+n$, 开始训练最终预测模型 (XGBoost- n)。最后一轮迭代的预测模型的损失函数需定义

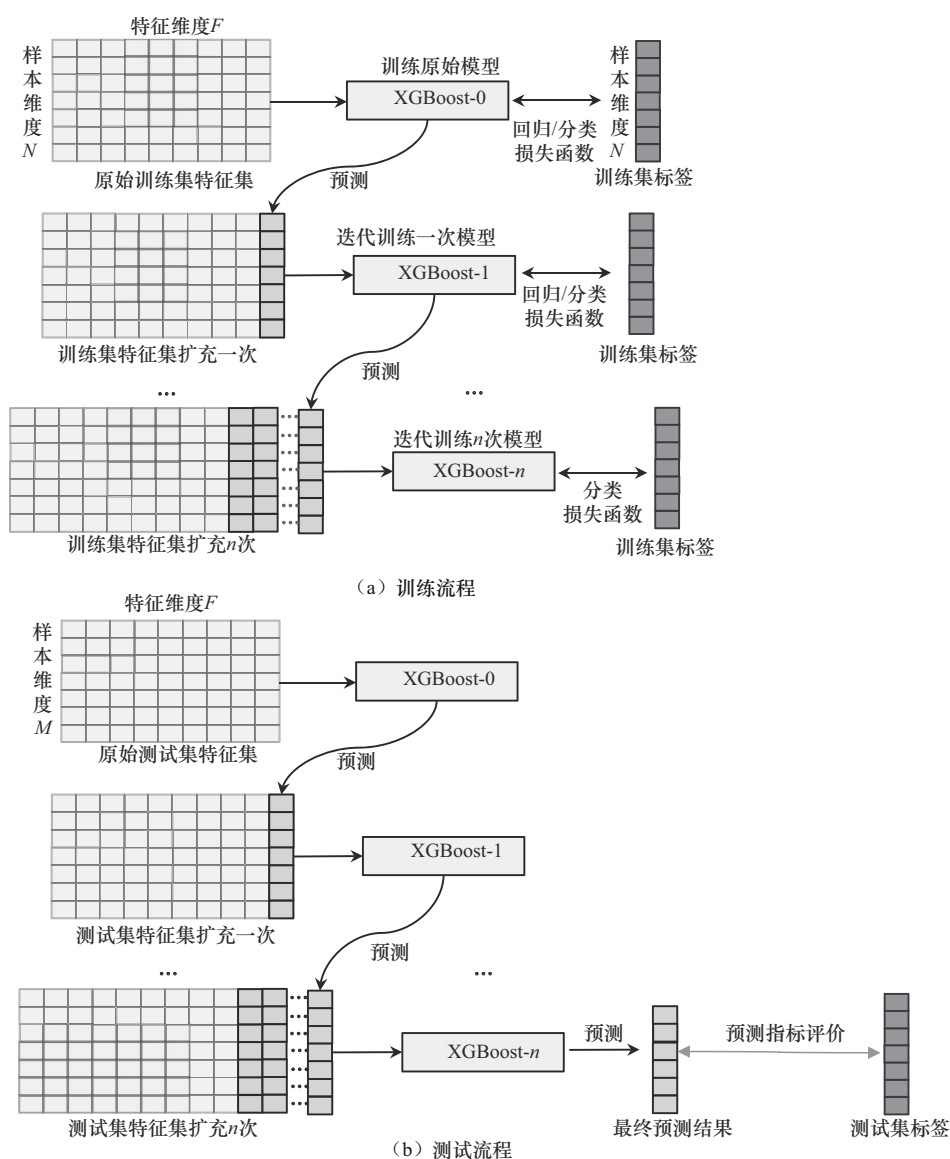


图2 RFA-XGBoost算法的训练与测试流程

为分类损失函数，用于用户产生投诉与否的直接判定。值得一提的是，除了最后一轮迭代的 XGBoost- n 模型为分类任务外，前序的 XGBoost 模型可以指定为分类任务，也可以指定为回归任务，实验部分将对两者的效果差异。

(6) 最终预测模型训练完成后，在扩充的测试集上开展测试，获取测试集用户的最终投诉预测结果。

综上，RFA-XGBoost 算法通过递归将模型的预测结果作为新特征添加到原始特征集中，有效扩展了特征空间。这种特征增强的方式能够捕捉到数据中更深层次的关联和模式，从而显著提升模型的预测性能。通过多次迭代，模型能够逐步学习到更加复杂和抽象的特征组合，从而更准确地描述数据。

4 实验分析

4.1 实验设置

XGBoost 的模型参数主要包括弱分类器数量、树模型的最大深度、学习率等，这些参数会影响模型的复杂度、训练速度和准确性。XGBoost 模型初始化的主要参数的含义、推荐值和部分候选值集合见表 3。

4.2 XGBoost 模型训练与评估

按照表 3 参数推荐值进行模型参数配置后，在训练集上开展 XGBoost 分类模型训练，并在测试集上进行效果评估。当数据集为包括业务域和运营域共 74 个指标时，基于 XGBoost 模型的测试集的混

淆矩阵如图 3 所示。本文所采用的模型效果评价指标主要包括准确率 (Accuracy)、召回率 (Recall)、精确率 (Precision) 和 F1 值 (F1-score)。

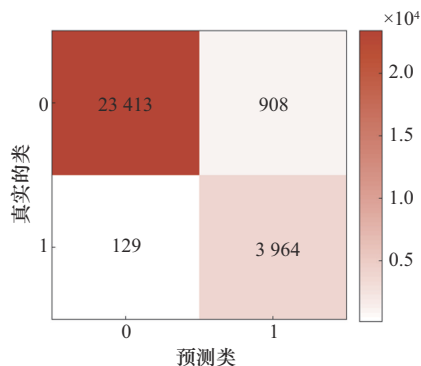


图 3 基于 XGBoost 模型的测试集的混淆矩阵

准确率表示分类正确的样本数量占总样本数量的比例，其数学表达式如下。

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (5)$$

其中，TP 为真阳性 (true positive)，指模型正确预测为正样本的样本数量；FP 为假阳性 (false positive)，指模型错误预测为正样本的样本数量；FN 为假阴性 (false negative)，指模型错误预测为负样本的样本数量；TN 为真阴性 (true negative)，指模型正确预测为负样本的样本数量。

召回率，指被正确分类的正样本数量占所有实际正样本数量的比例，其数学表达式如下：

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (6)$$

精确率，指被正确分类为正样本的数量占所有被预测为正样本的数量的比例，其数学表达式如下。

表 3 XGBoost 模型参数的含义、默认值和部分候选值集合

参数	推荐值	含义	部分候选值集合
learning_rate	0.3	学习率	{0.01, 0.02, 0.05, 0.1, 0.2, 0.3}
n_estimators	100	弱分类器数量	{50, 100, 150, 200, 300}
min_child_weight	20	叶子上的最小样本数	{1, 3, 5, 7, 9, 12, 15, 20}
max_depth	6	树模型的最大深度	{3, 4, 5, 6, 7}
subsample	1	样本随机降采样比例	{0.6, 0.7, 0.8, 0.9, 1}
colsample_bytree	1	特征随机降采样比例	{0.6, 0.7, 0.8, 0.9, 1}



$$\text{Precision} = \frac{TP}{TP + FP} \quad (7)$$

F1值是精确率和召回率的调和平均数,用于综合评估分类模型的性能,其数学表达式如下。

$$F1 = \frac{2\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (8)$$

选取3种不同的用户指标数据源——仅业务域26个指标、仅运营域48个指标和业务域+运营域74个指标——作为数据集,对比了基于XGBoost的移动网络潜在投诉用户预测效果。基于XGBoost模型的实验结果对比见表4。由表4可知,仅运营域48个指标作为数据集的模型预测平均准确率为94.60%,显著优于仅业务域26个指标作为数据集的模型预测平均准确率86.09%;融合了业务域+运营域共74个指标作为数据集的模型预测平均准确率为96.35%,显著优于仅运营域48个指标作为数据集的模型预测平均准确率。这验证了移动网络用户产生投诉是因为它受到多种客观和主观因素的共同影响,所以需要融合业务域和运营域的数据特征才能进行准确预测。因此,后续针对XGBoost和RFA-XGBoost模型的实验对比均在融合了业务域+运营域共74个指标数据源上开展。

4.3 XGBoost模型参数的影响

在融合了业务域和运营域共74个指标数据源上,进一步研究了包括弱分类器数量、树模型的最大深度、学习率等XGBoost模型参数对模型性能的影响,并开展了消融实验。

(1) 学习率

保持弱分类器数量为100,叶子上的最小样本数为20,树模型的最大深度为6,样本随机降采样比例为1,以及特征随机降采样比例为1。同时,从0.01开始逐步调整学习率,直至0.3,其

间将依次遍历{0.01, 0.02, 0.05, 0.1, 0.2, 0.3}这一候选值集合中的每一个值。在每个学习率下,都开展模型训练,并在训练完成后,使用测试集来评估模型的平均准确率,以便进行对比分析。XGBoost模型的测试集平均准确率随学习率的变化如图4所示。

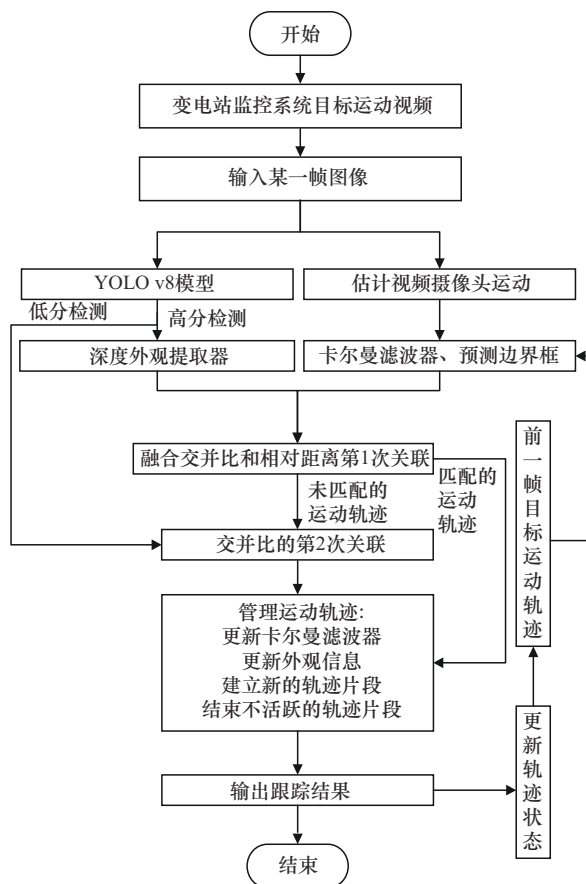


图4 XGBoost模型的测试集平均准确率随学习率的变化

从图4中可以看出,随着学习率的增加,模型的平均准确率呈现上升趋势。当学习率从0.01提升至0.1时,准确率提升显著,从92.70%提升至96.17%。当学习率进一步提升至0.2和0.3时,

表4 基于XGBoost模型的实验结果对比

算法	数据源	准确率	召回率	精确率	F1值
XGBoost	仅业务域26个指标	86.09%	63.11%	81.06%	66.54%
XGBoost	仅运营域48个指标	94.60%	85.42%	95.23%	89.39%
XGBoost	业务域+运营域74个指标	96.35%	90.40%	96.56%	93.13%

准确率稳定在约96.35%。

(2) 弱分类器数量

保持学习率为0.3，叶子上的最小样本数为20，树模型的最大深度为6，样本随机降采样比例为1，以及特征随机降采样比例为1。同时，从50开始逐步调整弱分类器数量，直至300，其间将依次遍历{50, 100, 150, 200, 300}这一候选值集合中的每一个值。XGBoost模型的测试集平均准确率随弱分类器数量的变化如图5所示。

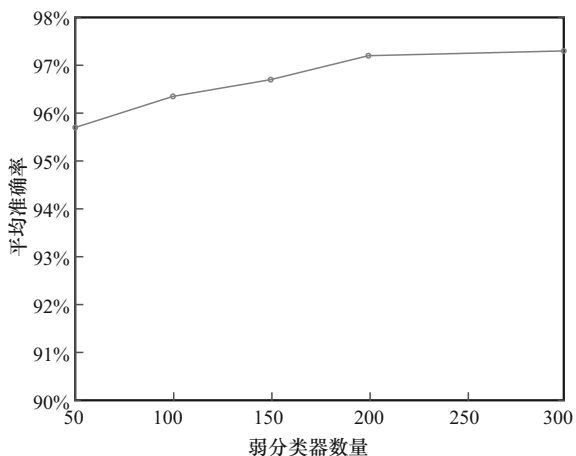


图5 XGBoost模型的测试集平均准确率随弱分类器数量的变化

从图5可以看出，随着弱分类器数量的增加，模型的平均准确率呈现不断上升的趋势。弱分类器数量越多，模型越能够捕捉到数据中的更多信息，从而提高预测的准确性，但是也会带来计算量大及过拟合的风险。

(3) 叶子上的最小样本数

保持学习率为0.3，弱分类器数量为100，树模型的最大深度为6，样本随机降采样比例为1，以及特征随机降采样比例为1。同时，从1开始逐步调整叶子上的最小样本数，直至20，其间将依次遍历{1, 3, 5, 7, 9, 12, 15, 20}这一候选值集合中的每一个值。XGBoost模型的测试集平均准确率随叶子上的最小样本数的变化如图6所示。

从图6可以看出，随着叶子上的最小样本数的增加，模型的平均准确率呈不断下降的趋势。

这是因为增加叶子节点的最小样本数限制会导致模型变得更加保守，降低了模型的复杂度，但也可能使得模型对数据的拟合能力降低。

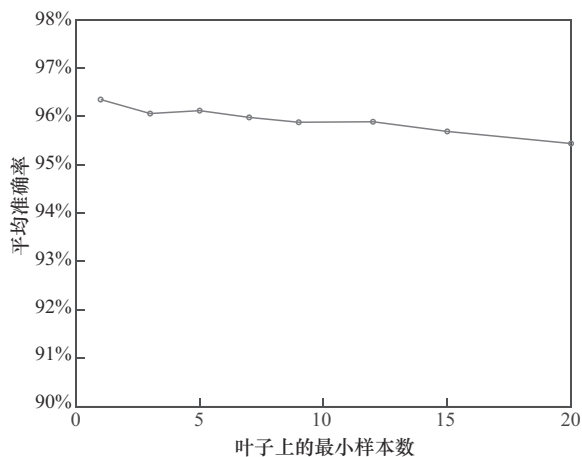


图6 XGBoost模型的测试集平均准确率随叶子上的最小样本数的变化

(4) 树模型的最大深度

保持学习率为0.3，弱分类器数量为100，叶子上的最小样本数为20，样本随机降采样比例为1，以及特征随机降采样比例为1。同时，从3开始逐步调整树模型的最大深度，直至7，其间将依次遍历{3, 4, 5, 6, 7}这一候选值集合中的每一个值。XGBoost模型的测试集平均准确率随树模型最大深度的变化如图7所示。

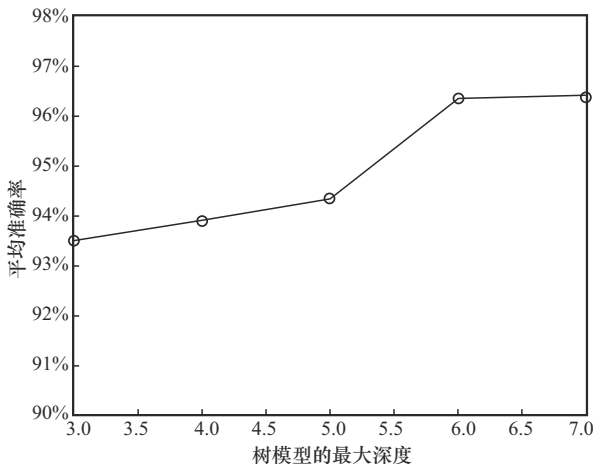


图7 XGBoost模型的测试集平均准确率随树模型最大深度的变化



从图7可以看出,随着树模型最大深度的增加,模型的平均准确率呈现不断增大的趋势。这是因为当树的最大深度增加时,模型能够学习到更加细致和复杂的特征划分,从而能够更准确地捕捉数据中的规律和模式,从而更好地拟合训练数据。

4.4 RFA-XGBoost模型训练与评估

针对RFA-XGBoost模型,继续按照表3参数推荐值进行模型参数配置后,在训练集上开展模型训练,并在测试集上进行效果评估。基于RFA-XGBoost模型实验结果对比见表5。此次实验采用了多种RFA-XGBoost方案,其中表5第5行表示经过1次XGBoost分类模型的实验结果,即第4.2节中的实验结果;表5第6行表示经过1次XGBoost分类模型训练后,将其在训练集上的预测结果作为新的特征,与原始74维训练集合并,形成新的75维特征训练集,重新训练XGBoost分类模型的实验结果;表5第7行表示经过1次XGBoost回归模型训练后,并将其在训练集上的预测结果作为新的特征,与原始74维训练集合并,形成新的75维特征训练集,重新训练XGBoost分类模型的实验结果;表5第8行表示经过XGBoost分类模型训练并将其在训练集上的预测结果作为新的特征,与原始训练集合并,不断形成新的特征训练集,重复10次,最终生成了84维特征训练集后重新训练XGBoost分类模型的实验

结果;表5第9行表示经过XGBoost回归模型训练并将其在训练集上的预测结果作为新的特征,与原始训练集合并,不断形成新的特征训练集,重复10次,最终生成了84维特征训练集后重新训练XGBoost分类模型的实验结果。

另外,我们在表5第1~4行分别补充了引言中提到的BP、GBDT、LightGBM以及Adaboost算法模型的实验结果。

对比表5第1~4行和第5行可以看出,BP、GBDT、LightGBM及Adaboost 4种机器学习算法的预测准确率均低于XGBoost。对比第6、7行和第5行可以看出,基于1次迭代的RFA-XGBoost相比原始XGBoost模型预测准确率有明显提升。同时,1次迭代时采用回归模型的平均准确率达98.29%,优于采用分类模型的平均准确率97.80%。对比第8、9行和第6、7行可以看出,相比基于1次迭代RFA-XGBoost模型,基于10次迭代的RFA-XGBoost预测准确率有明显提升。其中,采用回归模型的平均准确率从98.29%提升至98.76%,采用分类模型的平均准确率从97.80%提升至98.89%。可以看出,基于1次迭代的采用回归模型的RFA-XGBoost预测效果优于采用分类模型的RFA-XGBoost预测效果,然而基于10次迭代的采用回归模型的RFA-XGBoost预测效果反而不如采用分类模型的RFA-XGBoost预测效果。为了分析其原因或规律,本文对不同迭代

表5 基于RFA-XGBoost模型实验结果对比

算法	模型组合	准确率	召回率	精确率	F1值
BP		85.74%	61.87%	80.51%	65.00%
GBDT		90.17%	77.02%	85.76%	80.39%
LightGBM		92.42%	81.69%	90.24%	85.15%
Adaboost		89.18%	75.27%	83.47%	78.41%
XGBoost		96.35%	90.40%	96.56%	93.13%
RFA-XGBoost	1×XGBoost(cls)+ XGBoost	97.80%	94.44%	97.75%	96.00%
RFA-XGBoost	1×XGBoost(reg)+ XGBoost	98.29%	96.04%	97.87%	96.93%
RFA-XGBoost	10×XGBoost(cls)+ XGBoost	98.89%	97.47%	98.58%	98.02%
RFA-XGBoost	10×XGBoost(reg)+ XGBoost	98.76%	97.68%	97.93%	97.80%

数下的采用回归和分类模型的RFA-XGBoost预测平均准确率进行了对比。不同迭代数下的RFA-XGBoost预测平均准确率如图8所示。

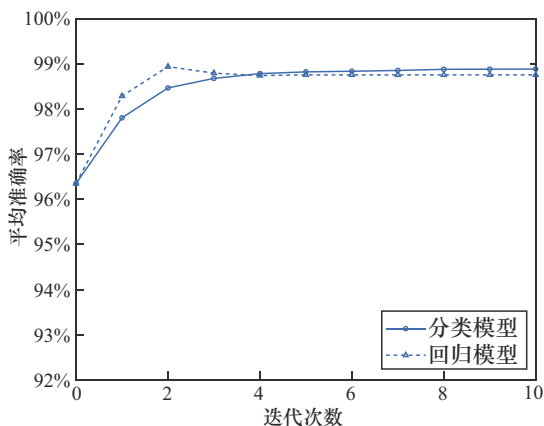


图8 不同迭代数下的RFA-XGBoost预测平均准确率

从图8可以看出，当迭代次数为0时，即原始XGBoost模型在测试集上的平均准确率为96.35%。当迭代次数大于0时，采用回归和分类模型的RFA-XGBoost预测平均准确率均有明显提升。其中基于回归模型的RFA-XGBoost预测平均准确率先升后降，最后收敛至98.76%；而基于分类模型的RFA-XGBoost预测平均准确率稳健提升，最终收敛至98.89%，略优于前者。此时，给出基于10次迭代的采用分类模型的RFA-XGBoost测试集的混淆矩阵如图9所示。从图9可以看出，相比图3中基于XGBoost模型的混淆矩阵，RFA-XGBoost模型能够有效提升识别准确率。

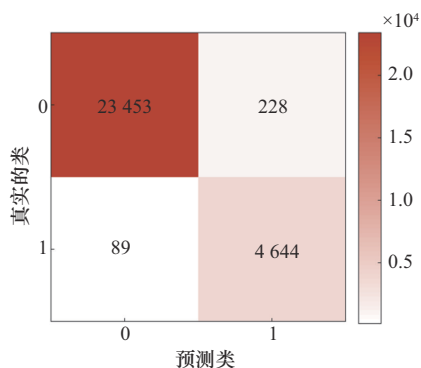


图9 基于10次迭代的采用分类模型的RFA-XGBoost测试集的混淆矩阵

综上，本研究通过对比分析，验证了RFA-XGBoost模型相较于XGBoost模型在移动网络投诉用户预测准确率上的显著提升。具体而言，无论是采用回归模型还是分类模型，RFA-XGBoost在引入迭代后均表现出预测性能的增强。值得注意的是，随着迭代次数的增加，基于回归模型的RFA-XGBoost预测准确率呈现出先增后稳的趋势，而基于分类模型的则持续稳健提升，并最终在10次迭代时略胜一筹。因此，在选择迭代次数时，需综合考虑预测精度的提升与计算效率的平衡。总体而言，RFA-XGBoost模型的引入为移动网络投诉用户的精准预测提供了有效手段。

5 结束语

本文成功地将多维数据分析应用于移动网络潜在投诉用户的预测中，并基于XGBoost构建了潜在投诉用户预测模型。通过采集移动网络用户广泛的业务域和运营域指标作为输入特征数据，在测试集上取得了96.35%的平均预测准确率。进一步地，本文创新性地提出了RFA-XGBoost预测模型，通过迭代新特征将平均预测准确率提升至98.89%。这一系列的研究成果，不仅验证了多维数据分析在移动网络潜在投诉用户预测中的有效性，也为移动网络运营商提供了有力的决策支持。

未来，将继续深化对移动网络用户行为模式的研究，挖掘更多潜在的特征数据，如用户的社交行为、位置信息等，以进一步提升预测模型的准确性和鲁棒性。这将有助于移动网络运营商更全面地理解用户需求，为用户提供更个性化的服务。

参考文献：

- [1] 胡怡. 消费保平台三大运营商投诉数据简报[J]. 消费电子, 2024(2): 17-20.



- HU Y. Complaint data briefing for the three major operators of consumer insurance platforms[J]. Consumer Electronics, 2024(2): 17-20.
- [2] 俞涛, 郝洁, 张小晖. 运营商用户投诉预判模型研究[J]. 移动通信, 2023(2): 208-210.
- YU T, HAO J, ZHANG X H. Research on the prediction model of operator customer complaints[J]. Mobile Information, 2023(2): 208-210.
- [3] 龙克树, 邓娟, 刘晓斌. 基于机器学习算法的运营商用户流失预判及应对策略研究[J]. 信息记录材料, 2020, 21(5): 202-205.
- LONG K S, DENG J, LIU X B. Research on predicting and coping strategies of operator's user loss based on machine learning algorithm[J]. Information Recording Materials, 2020, 21(5): 202-205.
- [4] 刘平. 移动网智能化网络投诉处理系统研究[J]. 通讯世界, 2021(7): 21-22.
- LIU P. Research on intelligent network complaint handling system of mobile network[J]. Telecom World, 2021(7): 21-22.
- [5] 吴亚楠. 客户无线网络投诉的预防措施研究[J]. 无线互联科技, 2020, 17(24): 15-16.
- WU Y N. Research on preventive measure of customer complaints in wireless network[J]. Wireless Internet Technology, 2020, 17(24): 15-16.
- [6] 万仁辉, 王洁, 戴鹏程, 等. 基于机器学习的网络投诉预测分析[J]. 电信工程技术与标准化, 2020, 33(8): 45-50.
- WAN R H, WANG J, DAI P C, et al. Prediction and analysis of network complaints based on machine learning[J]. Telecom Engineering Technics and Standardization, 2020, 33(8): 45-50.
- [7] 陈秀敏, 许向东, 黄毅华, 等. 基于数据挖掘的4G用户投诉预测[J]. 移动通信, 2017, 41(21): 30-36.
- CHEN X M, XU X D, HUANG Y H, et al. Prediction of 4G user complaints based on data mining[J]. Mobile Communications, 2017, 41(21): 30-36.
- [8] 宗宇雷, 梁童, 吴伟嘉, 等. 一种基于XGBoost算法的用户投诉预测方法[J]. 电信工程技术与标准化, 2023, 36(11): 7-12.
- ZONG Y L, LIANG T, WU W J, et al. A user complaint prediction method based on XGBoost algorithm[J]. Telecom Engineering Technics and Standardization, 2023, 36(11): 7-12.
- [9] CHEN T Q, GUESTIN C. XGBoost: a scalable tree boosting system[EB]. 2016, arXiv:1603.02754.
- [10] 王可争, 潘丹. 一种利用神经网络预测用户投诉的方法[J]. 通信世界, 2021(12): 45-48.
- WANG K Z, PAN D. A method for predicting user complaints by using neural network[J]. Communications World, 2021(12): 45-48.
- [11] 马洁. 基于数据挖掘与业务感知的投诉预测模型研究[J]. 电信工程技术与标准化, 2022, 35(S1): 115-121.
- MA J. Research on complaint prediction model based on data mining and business awareness[J]. Telecom Engineering Technics and Standardization, 2022, 35(S1): 115-121.
- [12] 周文杰, 严建峰, 杨璐. 基于深度学习的用户投诉预测模型研究[J]. 计算机应用研究, 2017, 34(5): 1428-1432.
- ZHOU W J, YAN J F, YANG L. Research on prediction model of complaint based on deep learning[J]. Application Research of Computers, 2017, 34(5): 1428-1432.
- [13] HINTON G E. Deep belief networks[J]. Scholarpedia, 2009, 4(5): 5947.
- [14] KE G L, MENG Q, FINLEY T, et al. LightGBM: A highly efficient gradient boosting decision tree[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems(NIPS'17). California: Curran Associates Inc, 2017: 3149-3157.
- [15] BEJA-BATTAIS P. Overview of AdaBoost: Reconciling its views to better understand its dynamics[EB]. arXiv preprint, 2023, arXiv:2310.18323.
- [16] FRIEDMAN J H. Greedy function approximation: A gradient boosting machine[J]. The Annals of Statistics, 2001, 29(5): 1189-1232.

[作者简介]



张鹏 (1980-), 女, 中国移动通信集团设计院有限公司湖北分公司工程师, 主要研究方向为网络研究、咨询及网络智能化。



高源 (1986-), 男, 中国移动通信集团设计院有限公司湖北分公司工程师, 主要研究方向为IT信息化。